Research on Intelligent Speech Signal Processing Using Vector Quantization(ベクトル量子化を用いた音声の知的情報処理に関する研究)

氏名	ハン シヒン 潘 志斌
授与学位	博士（工学）
学位授与年月日	平成 12 年 3 月 23 日
学位授与の根拠法規	学位規則第 4 条第 1 項
研究科，専攻の名称	東北大学大学院工学研究科（博士課程）電子工学専攻
学位論文題目	Research on intelligent speech signal processing using vector quantization （ベクトル量子化を用いた音声の知的情報処理に関する研究）
指導教官	東北大学教授 大見 忠弘
論文審査委員	主査 東北大学教授 大見 忠弘 東北大学教授 亀山 充隆 東北大学教授 川又 政征 東北大学助教授 小谷 光司

論文内容要旨

Chapter 1 Introduction

This chapter is an introduction to the whole work. Background, disciplines, previous work of intelligent speech signal processing is reviewed. Because speech is the most direct and vivid form to express people's thoughts and emotions by a sequence of pronunciations, intonations and so on, it is the first choice for human beings to communicate with each other. Furthermore, when Internet enters everyday life of ordinary people in late 90's, there appears the urgent need to fast, convenient, large-volume information exchange and processing method. Also, speech could be a most promising candidate. Suppose voice is taken as a medium for person's identity check to access administration system, speaker recognition system is essential. The goal of this work is to develop a speaker recognition system for field applications in noisy environment. By using this system, for the purpose of individual's access control, it is possible to recognize who you are among a population of several tens of registered people at high speed with acceptable success rate. Another goal of this work is to seek some potential applications of speaker recognition system, for example, the application of speech indexing to recorded audio source.

Chapter 2 Architecture of Speaker Recognition System

This chapter is about principles of speech signal processing and conventional architecture of speaker recognition system. In this chapter, the way of acquiring voice band speech, the way of pre-processing to enhance high frequency components of band limited speech signal is presented. Since speech signal is unstationary, it is generally divided into a series of frames with a length of 20 ms ~ 30 ms to get approximately stationary segments so that they can be treated in classic way such as FFT (e.g. short-time FFT) or spectrum estimation (e.g. linear prediction coding based methods, or LPC-based methods). To overcome discontinuity and spectrum leakage at the beginning and end of each frame, it must have joint parts between adjacent frames overlapped to some extent and windowed. For all frames, auto-correlation method is used to compute LPC coefficients and each set of LPC coefficients are then converted into cepstrum vector. This cepstrum vector is used as feature vector for further speaker recognition. Even though the content of speech is the same, depending on speaker and the way of speaking, the speech waveform and the extracted feature vector sequence from it are quite different. To compress and emphasis information included in speech, vector quantization (VQ) is used as a source coding method in this

work and speaker-dependent codebooks generated by Kohonen's LVQ are used as the templates for matching to classify and recognize speakers. To measure the similarity between input speech and codebooks, not only mean value μ of VQ distortion but also standard deviation σ' of it centered at μ and S/N after VQ are used to identify person. μ means how far the centroid of input speech is from that of each codebook and σ' gives to what extent the distribution of input speech is mismatched with that of each codebook. S/N describes relative distortion between input speech and each codebook in power. To decide which speaker the current piece of speech belongs to, decision by majority method is utilized among minimum value of μ, σ' and maximum value of S/N. Architecture of this speaker recognition system is consisted of 2 parts. One is for speaker registration and the other is for recognition of speaker.

Field experiment has been implemented for 58 persons in our lab. Each person is asked to read a piece of text 5 times to have about 10 second speeches in a computer room. People aging from teenage to 40s, foreigners coming from Italy, Korea, China and males or females are included. Japanese, English, Korean, Chinese pronounced by native speakers and foreigners are also included to check the variability of speech. With optimal learning conditions for codebook generation and compromised codebook size of 16 considering of time cost, for total 58 persons with 290 speeches, success rate can reach 92% in full matching rotation way.

Chapter 3 Noise Reduction by Exponential Average

This chapter is concerning with noise reduction technique. Success rate of speaker recognition is distinctively degrades due to effect of noise because noise makes the spectrum of clean speech corrupted depending on S/N ratio. Usually, to reduce effect of noise, SSB (spectrum subtraction) is used to minus noise spectrum from noisy speech spectrum in order to find out true speech spectrum in frequency domain. But this method has some problems such as how to estimate noise spectrum exactly from silent parts of noisy speech, how to guarantee the stability of LPC filter after spectrum subtraction, and how to reduce computation complexity. In this work, not in frequency domain but in time domain, by using nonlinear average, addictive noise can be reduced and cepstrum distortion due to noise can be compensated to some extent. Noisy speech signal in time domain changes rapidly. To follow this change, while doing average, it would be better to give more weights to recent samples and less to the samples in past so as to emphasis abrupt transient parts occurred lately. The weights could be an exponential function with the most recent sample weighted the heaviest and previous samples exponentially less. This algorithm can be implemented in one-step ahead average way recursively. The averaged output is computed through summing the current input sample multiplied by a smoothing factor k ($0<k<1$) and the summation of all previous input samples multiplied by $1-k$. Its computation complexity is rather low.

The noisy speech signals of 5 Japanese vowels "a, e, i, o, u" generated by adding white noise to the clean one with S/N ratio ranging from 0 dB to 20 dB are exponentially averaged to investigate the reduction effect of cepstrum distortion. Cepstrum distance between clean speech and noisy speech or exponentially averaged noisy speech is used as a measure of distortion. If cepstrum distance becomes smaller after exponential average, it means that exponential average can improve cepstrum distortion. For 5 Japanese vowels "a, e, i, o, u", cepstrum distortion can be improved by more than 5% relatively depending on S/N ratio. For noisy speeches with S/N of 0 dB ~ 20 dB, speaker identification among a population of 20 persons is also conducted. Success rate can be increased by 4% in the case of S/N is 20 dB.

Chapter 4 Implementation of Speaker Recognition System for Large Population

This chapter is about speaker recognition with large population. In practice, time consumption of recognition

for this case increases linearly with population registered in database. In the other way, the lager the codebook size is, the higher the success rate will be although success rate becomes saturated gradually when the codebook size reaches 16. Consequently, it is a better way to make the size of codebook as large as possible in order to have a higher success rate for speaker recognition. When both population and codebook size are rather large, it will become intolerable to wait for the result of recognition to appear in conventional way. In this chapter, 3 approaches are proposed to save the time for matching with codebooks in database or the time for registration of a codebook to database. 2-level hierarchical approach is through coarse-fine matching respectively so as to position some possible candidates firstly with small size codebook and then do high-resolution matching with lager size codebook. In this way, computation cost for matching can be reduced to about 7% for codebook size being 4 and 128 respectively and total speeches being 290 comparing with that of non- hierarchical approach and success rate keeps unchanged. Pre-learning approach is through making unknown input speech learn to let it compressed and to have its feature enhanced. Then it is used to do matching with codebooks in database in conventional way. This approach can reduce computation cost obviously for population larger than a threshold value and meanwhile make success rate improved. In the case of compressed speech being 32, codebook size being 16 and total speeches being 290, success rate of speaker recognition is improved by about 3%. On-line learning approach at registration phase is through making available sub speech learning while the rest of speech is being taken in. When the whole speech has been input, let it learn from beginning to end but with rather less times so as to incorporate correlations among all sub speeches into codebook. Because people feel that they begin waiting just from the ending of speech input or the starting of learning for the whole speech, it seems that the registration phase becomes quite fast, even though like in real time. In the case of sub speech being 1 second that will complete learning in real time, success rate saturates when times of learning for the whole speech is 20. Therefore, time for registration can be reduced to 16% in contrast to that in conventional way.

Chapter 5 An Application of Speaker Recognition Technique — Speech Indexing

This chapter is devoted to an application of speaker recognition technique that is called speech indexing. Speech indexing is a kind of search method for large amount of audio source based on information of speaker rather than contents of speech. To utilize speech data efficiently and effectively, in some cases, the location and distribution of designated speaker's voice or all speakers' voice in time domain should be known in advance. Through classifying a whole audio source based on speaker's information and attaching the results of speaker recognition in the form of speech index file, it will become easier and speedy to search and use this audio source in the future. A forum titled 『Forward to the international communications of "intelligence"』 recorded from an ordinary program broadcast by NHK education station on 6/11/99 is tested. There were totally 7 speakers (6 males and 1 female). Among them, one was American, one was Germany, and the others were Japanese (the foreigners were with interpreters). The speech to be classified is about 13 minus long totally. For the whole recording to be indexed, it should be divided into a series piece of speeches first. The trade-off length of piece is set to be 5 seconds based on pre-experiment. Exactness of indexing after correlation-based correction reaches 96%.

Chapter 6 Conclusions

This chapter is on principal results achieved through this work.

審査結果の要旨

ネットワーク技術の急速な発展に伴い、オンラインショッピングや電子商取引など、本人の認識、認証が極めて重要となるさまざまなサービスが提供されようとしている。本研究は、人間とコンピュータのインターフェイスとして最も自然な形態である音声に関して、ベクトル量子化を用いた知的音声情報処理により話者を高速にかつ高精度に認識する話者認識技術に関するものである。著者は、認識率を向上させるとともに、データベース登録話者数が増大した場合にも高速に処理が可能な新しい認識アルゴリズムを提案し、その有効性を実験的に明らかにした。本論文は、これらの成果を取りまとめたもので、全文６章よりなる。

第１章は序論である。

第２章は、一般的な話者認識アルゴリズムに関して述べており、線形予測符号化およびケプストラム分析による特徴量抽出機能と、ベクトル量子化による類似度検出および認識機能に関して概説するとともに、ソフトウェアにより構築した話者認識システムについて述べている。データベースに登録される個人データとなるコードブックは、Kohonen の自己組織化マップを用いた学習により生成しており、実験的に学習パラメータの最適化を行っている。

第３章は、音声信号における雑音除去手法について述べている。外乱信号源や伝送線路の特性等により入力信号である音声信号に重畳される雑音を除去することが、認識率を向上させるためにきわめて重要である。著者は、音声信号の雑音除去に初めて指数関数重み付け平均化手法を適用している。これは、従来の単なる移動平均化雑音除去手法と比べて雑音除去効果が高いだけでなく、再帰的演算により演算量も低減できる優れた手法である。意図的に雑音を重畳させた音声信号の話者認識実験により、指数関数重み付け平均化雑音除去により認識率が 91～92％であったものを４％以上向上することを実証している。これは実用上有効な成果である。

第４章は、登録話者数が増えた場合に増大するベクトル量子化演算量を低減し、効率的に話者認識を行う為の手法について述べている。まず、登録情報としての話者別特徴量コードブックを階層化し、次元数の少ないコードブックを用いて大まかな検索を行い候補を選択した後、次元数の大きい詳細なコードブックを用いた検索により話者を特定する階層化ベクトル量子化手法について述べている。さらに、入力となる音声信号からケプストラム分析により特徴量情報を抽出後、登録時と同様の学習により少数の典型的情報に圧縮してからベクトル量子化による認識を行う前学習処理手法を初めて提案している。ベクトル量子化演算の入力となるデータ量が減少することから、全体の演算量が 1/10 以下に低減できることを定量的に明らかにするとともに、この処理により、認識率も３％程度向上することを実験的に明らかにしている。これは、実用上きわめて重要な成果である。

第５章は、話者認識手法の応用例について述べている。具体的には、討論会や会議等の時間的に連続した複数話者の音声を自動的に話者毎に分類するシステムである。これは、連続音声を５秒程度のフレームに分割し、フレーム毎に前述の話者認識の手法を用いて誰の発声であるか識別してインデックス番号を付けるものである。発声の時間的相関性を利用した誤り訂正手法を考案して適用しており、約５％の認識率向上を実現している。

第６章は結論である。

以上要するに本論文は、話者認識において認識率を向上させ演算量を低減する新たなアルゴリズムを確立し、その効果を実証するとともに、複数話者による会議の話者認識システムを開発しており、電子情報工学の発展に寄与するところが少なくない。

よって、本論文は博士（工学）の学位論文として合格と認める。